SELECTION OF A BEST SPEECH RECOGNIZER FROM MULTIPLE SPEECH RECOGNIZERS USING PERFORMANCE PREDICTION

BACKGROUND

1. Field

This disclosure relates to speech recognition systems, more particularly to speech recognition systems using multiple speech recognizers with varying performance and operational characteristics.

2. Background

Speech recognition systems typically convert speech to text for dictation applications or to commands for command and control tasks. The speech is received through an incoming audio stream, converted and returned to the application as converted or recognized speech. The applications in use by the user generating the audio stream may include dictation systems, voice interfaces for menu driven applications, etc. The system may utilize cellular or landline phone systems, Voice-over-IP networks, multimedia computer systems, etc.

Speech recognition systems use modules referred to as speech recognizers, or recognizers, to perform the actual conversion. Performance in recognizers varies, even if recognizers are targeted at the same market. For example, a recognizer used in a system targeted to the dictation market from manufacturer A will perform differently than a recognizer targeted to the same market from manufacturer B. Additionally, 2 recognizers from a single manufacturer targeted at similar markets may perform differently. This occurs because different algorithms are used to perform the recognition and different speech models are used to drive the recognizers. Speech models may differ due to differing content and speakers used to generate the model, as well as the representation of this data.

Though current applications utilize only a single recognizer, more robust speech recognition systems may utilize several different recognizers. This allows the system to have

different recognizers available for different tasks and users. However, selecting the optimal speech recognizer for a given situation is problematic, as is a way to track and update performance records of the various recognizers in different situations, which allows better optimization.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention may be best understood by reading the disclosure with reference to the drawings, wherein:

Figure 1 shows one embodiment of a speech recognition system having a predictor module, in accordance with the invention.

Figure 2 shows one embodiment of a method of selecting a speech recognizer using performance prediction, in accordance with the invention.

DETAILED DESCRIPTION OF THE EMBODIMENTS

Figure 1 shows one embodiment of a speech recognition system 10 having a predictor and multiple recognizers. The speech recognition system receives an input audio stream through a port from some user of a speech recognition application at 12. The input audio stream is connected through some sort of connection 18 to the speech recognizers. The connection may be wireless, such as a cell phone connection, a local area network, or through the circuitry of a personal computer, as examples.

The speech recognizers may be partitioned across a network as in a distributed speech recognition system. In this case, the recognizers 14a-14n may process raw audio data or a more compact form generated by the front-end speech recognition processes that reside across the network 18. In the following discussions, we refer to the stream of data that the speech recognizers process as the "input stream" or "input audio stream" regardless if it is raw audio data or an intermediate form that may be generated by a distributed speech recognizer or any other form of compression. The invention is in no way limited by the specific architecture of the recognizers.

The recognizers, 14a-14n, receive the input stream. It must be noted that although it appears that there are 14 recognizers A-N, the use of the letter N is in the mathematical sense, where N is the number of recognizers decided upon by the system designer. In some

embodiments the predictor 22 may control the routing of the input stream to the recognizers. The recognizers receiving the input stream and converting it will be referred to as the enabled recognizers. The enabling of the recognizers by the predictor will be discussed in more detail below.

The enabled recognizers perform the speech recognition tasks. The output of the enabled recognizers would then be sent to an output switch 16. The predictor then selects a set of results from the results presented to the output switch. The basis of that selection is discussed in more detail below.

One embodiment of performing the selection of the appropriate recognizer is shown in Figure 2. As discussed above with regard to Figure 1, an input stream is received at 30. The input stream carries the speech input, as well other sounds and audio cues that may be used in deriving information pertinent to the speech recognition task.

For example, at 32, the input stream may be analyzed to determine characteristics of the communication channel. In some systems, analyzers exist that allow the system to determine the audio characteristics of the channel, for example, determining if the cellular or landline communication networks are in use. Other information may also be derived, including background noise and signal strength among others. Additionally, this analysis may determine characteristics of the communication device, for example determining if a speaker-phone or a wireless handset are in use This analysis may occur as part of deriving enabling information for the recognizers. Network-based information services such as CallerID in conjunction with a local or network-based database mapping calling number to channel and device characteristics may be utilized for similar effect.

Enabling information, as used here, is information used to select the recognizers that process the input stream. For example, the system may have some recognizers that are better for cell phone audio streams than other recognizers. Other characteristics may also determine

the recognizers that would be enabled for a particular input stream. Note that in some systems, the input stream may be routed to all recognizers even if they are not actively processing the input. This may occur, for example, if the recognizers are on a shared bus or communication network. All of the recognizers may be enabled, or only a few may be enabled, as will be discussed later. In some implementations, all recognizers may always be enabled at all times and in all situations.

For example, the characteristics of the device used to generate the audio stream or the communication channel in use may indicate a particular recognizer or recognizer type. The device could be a cell phone, even further defined as a particular brand of cell phone, a particular brand or type of landline phone, such as cordless, speakerphone or 'traditional' handset phone. Other examples exist including microphones and personal computers used in conjunction with phone lines or even over data networks, such as Voice-Over-Internet Protocol (VoIP phones). These are just examples of variations of the devices, and are not intended to limit scope of the invention in any way.

A more complex characteristic of the incoming stream is contextual information. Contextual information is that information related to the environment around the input stream, including characteristics of the user and information derived from the call using network services such as CallerID. This information may be obtained dynamically or may be predetermined.

As an example of dynamic information, the CallerID of a phone call that will have speech recognition performed on it may identify the call as international. Certain assumptions may be made about the quality of the audio stream based upon the identification of the call as international. An example of the predetermined information may be a user profile stored by the provider of the speech recognition application. The user may have a profile connected to their user identifier that identifies the user as an American male from the Deep South. This

may indicate that recognizers that have good performance for American men with southern accents be enabled.

In one embodiment of the invention, once the enabling information is received and those recognizers to receive the stream are enabled, the selection information is derived at 34. In an alternative embodiment, to be discussed later, the performance predictors or indicators may be updated and analyzed prior to the deriving the selection information. Deriving the enabling information allows the system to select the recognizers that will process the input stream. Deriving the selection information allows the system to select the best results from the recognizers.

The system may be configured so that it enables only a single recognizer determined to be the best for any given situation. In this case, the recognizers will return only one result, the results from that enabled recognizer. The system has derived enabling information; in this case, deriving selection information is trivial. Alternatively, the system may enable all recognizers and derive only the selection information, avoiding the enabling process and only using the predictor to select the results predicted to be the best. Finally, the system may use the predictor to do both the enabling and the selecting processes. The predictor would both enable some or all recognizers and the select the best results. In any of the above cases, the information used for either enabling, selecting or both may comprise the same general types of information discussed above.

In one embodiment, recognizers return their converted speech accompanied by one or more values that indicates the confidence the recognizer has in a particular result. We call these values individual-result confidence values. The predictor mechanism determines, for each recognizer in the system and for each situation, a recognizer-based confidence value. This recognizer-based confidence value is the predictor's estimation of the accuracy of each recognizer in a particular situation. The predictor tracks performance of individual

recognizers under a variety of conditions to determine recognizer-based confidence values. This includes correlating performance to channel and device characteristics, dialog state, user information and other contextual information. This allows the predictor to estimate, for example, the performance of recognizer A for user B communicating using a speakerphone.

The selection information will include at least one of the sets of information identified above: channel characteristics, device characteristics, user information, contextual information, dialog state, individual-result confidence values or the performance history of the particular recognizer. These sets of information will be referred to as performance-related information. Performance-related information exists in the predictor prior to receiving the converted streams from the recognizers or is derived from the input stream. This differs from the prior art, where the selection information is based only upon the converted streams and the active grammar. This prior art does not utilize performance-related information.

One element of performance-related information is a quantitative analysis of the costs of using a particular recognizer, in either financial or computational terms. Some manufacturers charge a licensing fee per use of their recognizers. This information may be factored into the performance related information to identify recognizers that may be too costly to be used in a particular context. Similarly, some recognizers use more computational resources than others. This factor may allow the system to select the best recognizer for the costs associated, either in licensing fees or load on the system.

The predictor selects the 'best' results based on the information available to it and tracks performance over time to improve the enabling and selection process. This prediction is important both for providing good system-level recognition accuracy and for optimizing resource usage by enabling only recognizers that may provide useful results. Information regarding the correctness of the recognition process is necessary to tune the predictor over time. We call this information "feedback" in the discussions that follow. Feedback may be

obtained directly through interaction with the user or by a variety of indirect means. We discuss this below.

The predictor gathers all information available and selects the results from the best recognizer, selecting a particular recognizer at 36. The information made available to the predictor in a particular embodiment will determine how the prediction is made in that embodiment. Several different kinds of information may be made available to the predictor for selection of a given recognizer. For example, the predictor may use recognizer-based confidence, analysis of the input stream, contextual information, and dialog states, as examples. Additionally, the predictor may store performance data derived either on-line or off-line to aid in prediction.

The measure of confidence may vary from recognizer to recognizer, so individual-result confidence values may have to be normalized to the same measurement scale. For example, one recognizer may provide a percentage out of 100% that indicates the recognizer's confidence that a particular result is accurate. Another may provide a rating on a scale from 1 to10. These would have to be converted to the same measure before they could be compared.

The individual-result confidence values may be used in a simple voting mechanism where several recognizers return a particular result. For example, the result may be "The quick brown fox." If 6 of the available recognizers return that particular result, that result will be given a higher confidence value than results that were returned only by one recognizer. This information can be leveraged with feedback and performance history, as will be discussed further. Note that individual-result confidence values are not necessary to implement a voting mechanism, nor are they required for implementation of these feedback and performance tracking mechanisms.

In one embodiment of the invention, the flow shown in Figure 2 changes to include 40 and 42. The input stream is routed to the enabled recognizers in 40. The enabled recognizer provides results and associated recognizer information that may include individual result confidence values, for example, to the predictor at 42. The predictor then uses the recognizer information as part of the selection information in 34.

As mentioned above, the analysis of the input stream may identify certain channel or device characteristics that can be used in the selection information. Certain recognizers may be provided by a manufacturer and targeted for a particular channel or device. This information can be used to weight the information received from recognizer, including individual result confidence values.

Similarly, contextual information, as discussed, above can also be used to predict the best recognizer. Contextual information may include gender, age, ethnicity, whether the speaker speaks the language of the recognizers as a first language, among other personal information about the user. Also, the channel and device characteristics may be included in the contextual information.

The channel and device characteristics were discussed above as part of an analysis of the input stream. It must be noted that even systems that do not perform this signal-processing analysis may be able to use device and channel characteristics that are known or implied from the context of the call, rather than from an analysis of the input stream. For example, the CallerID, along with local or network based databases associating calling number with device characteristics, may identify the incoming call as being from a cellular phone. While all of these different sources of information may be used together, they may also be used alone, and systems not performing input stream analysis may still have channel and device characteristics available as part of the selection information.

Another source of information that can be factored into the selection information is the dialog state of the interaction between the user and the application requesting speech recognition. The application knows the state of the dialog and can provide that information to the speech recognition system. The speech recognition system can use that information to predict the best recognizer. For example, certain recognizers may be better for control keywords. A user calls into the system and navigates the menus using control keywords and then starts a dictation process. The speech recognition system may have knowledge about which recognizers are optimized for the control keywords and which are best for the dictation. It can then use this knowledge in the selection information. Additionally, a variety of recognizers optimized for dictation may be available, for example. However, they perform differently in different domains, with one being optimized for legal use, one optimized for medical use, and a third for general use for example. If the system knows that the user is dictating a legal memo based on the current state of the dialog, it may use the legal-dictation-optimized recognizer. Alternatively, the dialog state may indicate that the more general dictation recognizer is more appropriate. This knowledge may be used in the selection and enabling processes.

In the examples above, the speech recognition system may change recognizers for a given user. The streams entering the system initially are control keyword audio streams. At some point, the input streams may change to dictation audio streams. At that point, the system may select a different recognizer. This system is dynamic and flexible.

For example, assume there are two recognizers. Recognizer A does speaker and channel adaptation and adapts slowly but ends up being very accurate if it has enough time to adapt. Recognizer B performs moderately well but requires no adaptation time. Initially, the system chooses B, but allows both to process the audio. This allows A to adapt. Eventually, Recognizer A has adapted sufficiently to cause the system to expect that it will out perform

the Recognizer B. The recognizers may then provide this information to the predictor. At this point, the predictor will use results from Recognizer A. With regard to enabling the recognizers, if applicable, Recognizer A is being enabled based upon its expected future performance, rather than its current performance.

Generally, the selection information with regard to Recognizer A will be updated, and the updated selection information will cause a different recognizer to be selected than was used in previous interactions. An interaction could be an utterance within a session. For example, the user makes several utterances separated by silence. For the first few utterances, the system may use Recognizer B, as Recognizer A adapts. For the remaining utterances, the system may switch to Recognizer A, not the recognizer used for the previous utterances. Alternatively, an interaction could be an entire session where the user enters a speech recognition process and then ends it, such as dialing into a telephone-based system, and ending by hanging up. . In this example, Recognizer B may be used for a first few sessions with a particular user, while Recognizer A is adapting, and then the system may select Recognizer A after it has adapted.

Up to this point in the process, the knowledge of the recognizers is preexisting knowledge, probably provided when the recognizer was initially received or installed. The predictor uses all of the information available to predict the best results by identifying the best recognizer for a given situation. These results are then returned to the application. All of the above measures and information can be updated dynamically using feedback mechanisms.

The predictor may utilize data on past performance to perform the enabling and selection processes. As discussed above, we call this data "feedback". Feedback may be explicit or implicit as is demonstrated below. These feedback mechanisms are only intended as examples and are not intended to limit scope of the invention in any way.

Referring to Figure 2, after the results are returned to the application, feedback may be generated at 44, as an option. This feedback may then be used to update information about the recognizers at 46, such as their performance with a given characteristics in the contextual information, a weighting that is applied to their individual-result confidence values, their optimization for various channels and devices, among many others. This is optional. This updated information may also be used when deriving the enabling information or the selection information, or both. The feedback may be generated in several ways, and at several different times. For example, feedback generated off-line will not occur at the place in the process shown here, but much later. In some systems, the user may provide the feedback, directly indicating the recognition accuracy. This may occur, for example, in dictation applications. The predictor may track the user's corrections and use that information to tune the prediction process. In many applications, this is too cumbersome to be practical. A variety of indirect measurement methods are discussed below.

In some speech recognition systems, the feedback may be from human evaluation. After results are returned from the enabled recognizers, they may be stored with the original audio stream for off-line evaluation. A human could listen to the audio stream and then review the results provided and make indications as to which enabled recognizers were correct. This evaluation would then be coded and stored for later access by the predictor.

Another possibility for feedback generation is the voting based estimation mentioned above. Similar to the pure voting scheme used to select a particular result, measurements for each recognizer may be maintained. If the recognizer was one that returned the results with the highest number of votes, a recognizer-based confidence value may be raised.

The voting scheme could also be leveraged to weight results returned from various recognizers. For example, a recognizer with a recognizer-based confidence value of 90% returns a result with an associated individual result confidence value of 95%. The weighted

individual-result confidence value for that result would then be (0.90)*(.95) or 0.86. Another recognizer with a recognizer-based confidence value of 99% may return a particular result with an associated individual result confidence value of 90%. The weighted individual-result confidence value for that result is (0.99)*(0.90) or 89%. In this manner, even though the first recognizer returned an individual-result confidence value for the results higher than the second recognizer, the second recognizer's results would 'win.' This is because the second recognizer has performed correctly in the past 99% of the time.

Prior speech recognition systems that have utilized multiple recognizers have not attempted to track performance information over time. Rather, they implement a simple voting scheme that evaluates each utterance independent of previous recognizer performance. This is discussed in Barry, T.; et al., "The simultaneous use of three machine speech recognition systems to increase recognition accuracy," Aerospace and Electronics Conference, 1994. NAECON 1994, Proceedings of the IEEE 1994 National Page(s): 667 - 671 vol.2. The invention disclosed here uses performance histories and extensive contextual information to make the multiple speech recognizer system more robust and better performing.

Implementation of the invention is left to the system designer. The speech recognition system of Figure 1 may be built from scratch to include the predictor. Alternatively, the predictor may be included in existing speech recognition systems as an upgrade. While it is possible that the predictor may be a dedicated processor or application-specific integrated circuit (ASIC), it may also be embodied in software loaded into the speech recognition system. In this instance, the software would perform the steps of the invention.

The software may be loaded into a digital signal or general-purpose processor via some form of computer readable medium, or as an image file into a digital signal processor. In either example, as well as others, the invention would be contained on an article including

machine-readable code, where execution of that code would implement the methods of the invention.

Thus, although there has been described to this point a particular embodiment for a method and apparatus for selecting a best speech recognizer from multiple speech recognizers, it is not intended that such specific references be considered as limitations upon the scope of this invention except in-so-far as set forth in the following claims.